

# Classifier comparison using precision

**Lovedeep Gondara**

School of Computing Science  
Simon Fraser University  
lgondara@sfu.ca

## Abstract

New proposed models are often compared to state-of-the-art using statistical significance testing. Literature is scarce for classifier comparison using metrics other than accuracy. We present a survey of statistical methods that can be used for classifier comparison using precision, accounting for inter-precision correlation arising from use of same dataset. Comparisons are made using per-class precision and methods presented to test global null hypothesis of an overall model comparison. Comparisons are extended to multiple multi-class classifiers and to models using cross validation or its variants. Partial Bayesian update to precision is introduced when population prevalence of a class is known. Applications to compare deep architectures are studied.

## Introduction

Classification models are often compared to test a global null hypothesis ( $H_0$ ) of one performing significantly better than other(s). There is no standard framework, nor it is clearly defined what statistical tests to use on what performance metrics. This often results in arbitrary choices (Demšar 2006) with classification accuracy being used most often.

Precision is an important performance statistic, especially useful in rare class predictions. It is the probability of positive prediction conditioned on classifier results and is often calculated per-class with an average reported as a point estimate

$$P_C = \frac{1}{N} \sum_{i=1}^N P_{Ci} \quad (1)$$

where  $P_{Ci}$  is precision for class  $i$  and  $N$  is number of classes.

Ideally, we can then compare  $\hat{P}_C$  for different models. However, one major issue is the use of same dataset to build multiple models, which results in correlated precision values. If not accounted for, it will result in biased inference. Another issue is the use of statistical tests designed to compare proportions, precision being a conditional probability is inherently different. It is also often overlooked that some classes might be of greater interest compared to others, thus needing higher precision scores. Reporting and comparing average precision will dilute this effect. An example is of classifying a malignant tumor from others, where we would

prefer a model with higher precision for malignant class. Second example can be of desired higher precision to identify stop signs and pedestrians for an autonomous vehicle.

As McNemar’s test (McNemar 1947), paired t-test and Wilcoxon’s signed rank test can be used to compare classifiers based on their sensitivity, specificity and accuracy or mean of it. There is a need of appropriate statistical test that can be used to compare models using precision. As much time and effort goes into model building, model comparison should be treated the same. To best of our knowledge, there is no present literature in machine learning that reviews or introduces any tests to compare correlated precision values. However, similar studies are present for other metrics (Aslan, Yıldız, and Alpaydın ; Benavoli et al. 2014; Demšar 2006; Dietterich 1998; Joshi 2002; Nadeau and Bengio 2003).

In this paper we present survey of statistical methods that can be used to compare classifiers based on precision. Comparisons are made on per-class bases, with methods provided to combine inference for an overall classifier comparison. Methods are introduced to compare classifiers in cross validation (single  $k$ -fold and  $n$  times  $k$ -fold) settings commonly used by practitioners. We show that these methods can be used for simultaneous comparison of multi-class multiple classifiers. We also present a partial Bayesian approach to update precision when class prevalence is known and demonstrate application of these methods to compare models based on deep architectures. We intend to enrich machine learning literature by providing methods to be used for model comparison using precision. Methods presented are not intended to replace or compete with existing statistically sound methods, but to supplement them. All methods presented in this paper can also be used to compare recall values.

Next section presents an overview of statistical methods followed with empirical evaluation and our extensions. Final section presents our conclusions and recommendations.

## Statistical tests

### Standard notation

We start by defining precision. Given two binary classifiers  $C_1$  and  $C_2$ , we can write the results as Table 1 and Table 2.

Estimated precision for classifier  $C_1$  and  $C_2$  can be writ-

Table 1: True label vs predicted label for  $C_1$ .

$C_1$ PREDICTED LABEL				
	1	0		TOTAL
TRUE LABEL	1	A	B	$T_4$
	0	C	D	$T_3$
	TOTAL	$T_1$	$T_2$	

Table 2: True label vs predicted label for  $C_2$ .

$C_2$ PREDICTED LABEL				
	1	0		TOTAL
TRUE LABEL	1	E	F	$T_8$
	0	G	H	$T_7$
	TOTAL	$T_5$	$T_6$	

ten as

$$\hat{P}_{C1} = \frac{A}{A+C}, \hat{P}_{C2} = \frac{E}{E+G} \quad (2)$$

It is clear from (2) that precision is the probability of making correct predictions conditioned on classifier predicted label, i.e.  $Pr(True - label | Predicted - label)$ .

Our null hypothesis for comparing  $\hat{P}_{C1}$  and  $\hat{P}_{C2}$  is

$$H_0 : \hat{P}_{C1} = \hat{P}_{C2} \quad (3)$$

i.e. "For a given training set, the estimated precision for classifiers  $C_1$  and  $C_2$  is not statistically significantly different."

But as same dataset is used to calculate  $\hat{P}_{C1}$  and  $\hat{P}_{C2}$ , values compared are not independent. Also being a conditional probability, precision does not lend itself well for methods designed for proportions. A suitable test that takes these factors into account is needed. Several such tests are presented in following sections.

### Methods based on marginal regression framework

Let  $D_i$  be true label where 1 is the class label of interest and 0 otherwise,  $x_i$  is the predicted label by classifier and  $f_i$  is the classifier used (1 for  $C_1$  and 0 for  $C_2$ ). We can define  $P_{C1}$  and  $P_{C2}$  as

$$P_{C1} = Pr(D_i = 1 | f_i = 1, x_i = 1) \quad (4)$$

$$P_{C2} = Pr(D_i = 1 | f_i = 0, x_i = 1) \quad (5)$$

(Leisenring, Alono, and Pepe 2000) used methods based on Generalized Estimating Equations (GEE) (Liang and Zeger 1986) by restructuring the performance data of a medical diagnostic test to fit a marginal regression framework. In our case, restructured data will have one row per classifier prediction. For a two classifier comparison we will have two rows per observation. Implementation details are given in Algorithm 1.

As the dataset has repeated observations (multiple observations per data point), we use GEE based Generalized Linear Model (GLM). Parameter estimation and associated standard errors are calculated using robust sandwich variance estimates (Huber 1967; White 1980; Liang and Zeger 1986).

Using logit link function we can fit the model

$$Pr(D_i = 1 | f_i, x_i = 1) = \frac{\exp(\alpha + \beta f_i)}{1 + \exp(\alpha + \beta f_i)} \quad (6)$$

---

**Algorithm 1** GEE based comparison using logistic regression

---

**Require:**

- 1: k: Number of classes
  - 2: c: number of classifiers to be compared
  - 3: y: outcome/target class
  - 4:  $\hat{y}$ : predicted class
  - 5:
  - 6: Generate *id* variable for each observation
  - 7:
  - 8: **for** classifier in 1:c **do**
  - 9:     Save prediction as a
  - 10:    Save classifier name and y as b
  - 11:    Merge a and b horizontally
  - 12: **end for**
  - 13: Stack c datasets vertically to generate a single dataset d
  - 14: **for** i in 1:k **do**
  - 15:     Subset dataset d with  $\hat{y} = i$
  - 16:     Fit a GEE based GLM with binomial link and independent working correlation matrix as shown in (6)
  - 17:     Save required parameters
  - 18: **end for**
- 

For a two classifier comparison  $\exp(\beta)$  is the ratio of true prediction given classifier predicted value from  $C_2$  vs  $C_1$ . It describes the degree to which one classifier is more predictive of true classification than other.

Advantage of this method is the possibility of simultaneous comparison of multiple classifiers over multiple datasets. Model estimates such as odds ratio and related confidence intervals can be calculated for supplemental information.

**Empirical Wald test** From model (6), Wald test for null hypothesis can be used to test  $H_0 : \hat{\beta} = 0$ . Which is equivalent to  $H_0 : \hat{P}_{C1} = \hat{P}_{C2}$ , given as

$$\hat{T}_W = \frac{\hat{\beta}^2}{V(\hat{\beta})} \quad (7)$$

where denominator is second(last) diagonal element of empirical variance-covariance matrix.

Reformulation of empirical Wald test statistic was given in (Kosinski 2013) as

$$\hat{T}_W = \frac{\hat{\beta}^2 \times \hat{P}_{C1}(1 - \hat{P}_{C1})\hat{P}_{C2}(1 - \hat{P}_{C2})}{\frac{\hat{P}_{C1}(1 - \hat{P}_{C1})}{T_1} + \frac{\hat{P}_{C2}(1 - \hat{P}_{C2})}{T_5} - 2\hat{C}_P(\frac{1}{T_1} + \frac{1}{T_5})} \quad (8)$$

where  $\hat{\beta}$  is the estimated regression parameter from (6).

It is to be noted that GEE based Wald test statistic is similar to multinomial Wald statistic as  $\hat{\beta} = \text{logit}\hat{P}_{C1} - \text{logit}\hat{P}_{C2}$ .

**Score test** Score test statistic based on GEE was given in (Leisenring, Alono, and Pepe 2000) as

$$\hat{T}_S = \frac{\{\sum_{i=1}^N D_i(T_i - m_i \bar{Z})\}^2}{\sum_{i=1}^N (D_i - \bar{D})^2 (T_i - m_i \bar{Z})^2} \quad (9)$$

where  $T_i = \sum_{j=1}^{m_i} Z_{ij}$  is number of positive predicted labels for observation  $i$ . In a two classifier setting,  $T_i$  is the indicator variable for correct predictions. Also

$$\bar{D} = \frac{(\sum_{i=1}^N \sum_{j=1}^{m_i} D_{ij})}{\sum_{i=1}^N m_i} \quad (10)$$

and

$$\bar{Z} = \frac{(\sum_{i=1}^N \sum_{j=1}^{m_i} Z_{ij})}{\sum_{i=1}^N m_i} \quad (11)$$

$N$  is the number of observations with at least one true predicted label and  $m_i$  is number of true predicted labels for  $i^{th}$  observation.

Simple reformulation of (9) was given by (Kosinski 2013)

$$\hat{T}_S = \frac{(\hat{P}_{C1} - \hat{P}_{C2})^2}{\{(\hat{P}_{CP}(1 - \hat{P}_{CP}) + \hat{W}_P - 2\hat{C}_P)(\frac{1}{T_1} + \frac{1}{T_5})\}} \quad (12)$$

where

$$\hat{W}_P = (2\hat{P}_{CP} - \hat{P}_{C1} - \hat{P}_{C2})(2\hat{P}_{CP} - 1) \quad (13)$$

$\hat{P}_{CP}$  is pooled precision, estimated from Table 1 and 2 as

$$\frac{A + E}{A + C + E + G}. \quad (14)$$

## Methods based on relative precision

Table 1 and 2 is restructured as Table 3 estimated precision

Table 3: True label vs predicted label for  $C_1$  and  $C_2$

	TRUE LABEL=0		TRUE LABEL=1	
	$C_2=1$	$C_2=0$	$C_2=1$	$C_2=0$
$C_1=1$	$n_1$	$n_2$	$n_5$	$n_6$
$C_1=0$	$n_3$	$n_4$	$n_7$	$n_8$

for  $C_1$  and  $C_2$  is now rewritten as

$$\hat{P}_{C1} = (n_5 + n_6)/(n_1 + n_2 + n_5 + n_6) \quad (15)$$

$$\hat{P}_{C2} = (n_5 + n_7)/(n_1 + n_3 + n_5 + n_7) \quad (16)$$

Relative precision (RP)  $\hat{P}_R$  is defined as

$$\hat{P}_R = \frac{\hat{P}_{C1}}{\hat{P}_{C2}} \quad (17)$$

Using log transformation, variance of  $\log \hat{P}_R$  is estimated with  $\hat{\sigma}_P^2/N$ . Where

$$\hat{\sigma}_P^2 = \frac{1}{(n_5 + n_7)(n_5 + n_6)} \times \{n_6(1 - \hat{P}_{C2}) + n_5(\hat{P}_{C2} - \hat{P}_{C1}) + 2(n_7 + n_3)\hat{P}_{C1} \times \hat{P}_{C2} + n_7(1 - 3\hat{P}_{C1})\} \quad (18)$$

100(1 -  $\alpha$ )% confidence intervals are then constructed as

$$\log \hat{P}_R \pm Z_{1-\alpha/2} \sqrt{\frac{\hat{\sigma}_P^2}{N}} \quad (19)$$

19 is exponentiated to obtain upper and lower limits of  $\hat{P}_R$ . Confidence intervals from (19) can be used to test

$$\hat{P}_R = \xi \quad (20)$$

where

$$H_0 : \hat{P}_{C1} = \hat{P}_{C2} \iff \xi = 1 \quad (21)$$

$H_0$  is rejected if lower confidence interval of  $\hat{P}_R$  is greater than  $\xi$  or upper confidence interval is less than  $\xi$ .

## Empirical evaluation

### Experimental setup

To demonstrate feasibility of methods described in this paper, we used publically available datasets from UCI machine learning repository (Blake and Merz 1998) with varying characteristics and sample sizes as shown in Table 4.

Table 4: Datasets used for evaluation

DATASET	INSTANCES	ATTRIBUTES	CLASS
WILT	4889	6	2
DIABETIC RETINOPATHY	1151	20	2
PHISHING	2456	30	2
BANK NOTE	1372	5	2
MAGIC	19020	11	2
URBAN LAND COVER	675	148	9

If a dataset was not already partitioned, training-test split of 70%-30% was used. Although most evaluations were performed using fixed training and test splits, same procedures can be adapted when using cross validation as shown in following sections.

Random forest (Breiman 2001) with 1000 trees and Naive Bayes were used for initial comparisons. All comparisons were implemented in R (R Core Team 2015; Halekoh, Højsgaard, and Yan 2006; Stock and Hielscher 2013; Hongying Dai and Cui 2014). Sample code is provided as supplemental material.

### Comparison using Generalized Score and Empirical Wald test

Per-class precision was calculated for all datasets. Values were then compared using Generalized Score (GS) and Wald test statistic (GW). Results are shown in Table 5. Results from GS and GW statistics agree for all comparisons. GS has more power and performs better with small sample size compared to GW (Leisenring, Alono, and Pepe 2000).

### Comparison based on relative precision

Concerns around use/misuse of p-values (Nickerson 2000; Gill 1999; Anderson, Burnham, and Thompson 2000) can be alleviated by reporting RP and related confidence intervals (CIs). Although if necessary, p-value and a test statistic

Table 5: Comparison of  $C_1$  and  $C_2$  using GS and GW

DATASET	CLASS	NB	RF	P-GS	P-GW
WILT	N	0.65	0.74	<0.0001	<0.0001
	W	0.73	0.98	0.001	0.003
DIAB. RET.	0	0.54	0.63	0.002	0.0002
	1	0.76	0.67	0.07	0.09
PHISHING	-1	0.95	0.98	<0.0001	<0.0001
	1	0.92	0.96	<0.0001	<0.0001
BANK NOTE	0	0.83	0.99	<0.0001	0.0001
	1	0.85	0.98	<0.0001	<0.0001
MAGIC	G	0.72	0.88	<0.0001	<0.0001
	H	0.70	0.87	<0.0001	<0.0001
LAND COVER	ASP	0.94	0.95	0.87	0.87
	BLD	0.91	0.85	0.04	0.04
	CR	0.77	0.69	0.25	0.26
	CNR	0.85	0.78	0.04	0.05
	GRS	0.76	0.75	0.77	0.77
	PL	0.92	0.92	0.95	0.95
	SHD	0.82	0.79	0.48	0.48
	SL	0.36	0.60	0.01	0.02
	TR	0.70	0.86	0.0001	0.001

NB: Naive Bayes (Precision), RF: Random forest (Precision), P-GS: P-value from GS, GW: P-value from GS.

can be calculated as well. Comparison results using RP are shown in Table 6. Results using RP are in agreement with GW and GS. Standard statistical interpretation can be used with CIs. CIs not including '1' indicate a statistical significant difference.

Another advantage of using RP is the nice graphical representation of results it lends itself to. An example of this is shown in Figure 1 with results plotted from Table 6 for first five datasets using a forest plot. Box represents point estimate with extended lines representing 95% CIs. Reference line at '1' is plotted for visual inspection of a statistical significant difference. Confidence intervals not overlapping the reference line are considered significant.

### Combining inference

Investigators are often interested in testing a global  $H_0$  of an overall classifier comparison. Methods presented in this paper provide a per-class granular control, an overall comparison is still desired. Results in Table 5 and Table 6 replicate a multiple comparison scenario, which if not accounted for can pose a challenge to control classifier wide Type I error rate. Common methods to adjust for multiple comparisons include Family Wise Error Rate (FWER) correction or controlling for False Discovery Rate (FDR). But, in this case we also need to acknowledge dependence between p-values, resulting from probable contribution of observations across classes. Hence, specially tailored methods to combine dependent p-values are needed.

For dependent p-values, the distribution of combined test statistic does not have an explicit analytical form. It is approximated using a scaled version (Li, Williams, and Cui 2011) with a new  $\chi^2$  distribution. Satterthwaite method (Satterthwaite 1946) was used in (Hongying Dai and Cui 2014)

Table 6: Comparison of  $C_1$  and  $C_2$  using Relative precision

DATASET	CLASS	RP(95% CI)	P-VALUE
WILT	N	0.88 (0.85,0.92)	<0.0001
	W	0.75 (0.62,0.91)	0.003
DIAB. RET.	0	0.85 (0.78,0.92)	0.0001
	1	1.13 (0.99,1.29)	0.06
PHISHING	-1	0.97 (0.96,0.98)	<0.0001
	1	0.96 (0.95,0.97)	<0.0001
BANK NOTE	0	0.83 (0.79,0.88)	<0.0001
	1	0.87 (0.82,0.92)	<0.0001
MAGIC	G	0.82 (0.81,0.83)	<0.0001
	H	0.80 (0.77,0.83)	<0.0001
LAND COVER	ASP	0.99 (0.92,1.10)	0.87
	BLS	1.07 (1.0,1.15)	0.04
	CR	1.12 (0.92,1.35)	0.26
	CNR	1.1 (1.0,1.17)	0.05
	GRS	1.02 (0.90,1.16)	0.77
	PL	1.01 (0.81,1.3)	0.95
	SHD	1.03 (0.94,1.14)	0.48
	SL	0.6 (0.39,0.93)	0.02
	TR	0.81 (0.73,0.91)	0.0002

RP: Relative precision, 95% CI: Confidence intervals, comparisons are based on "NB/RF"

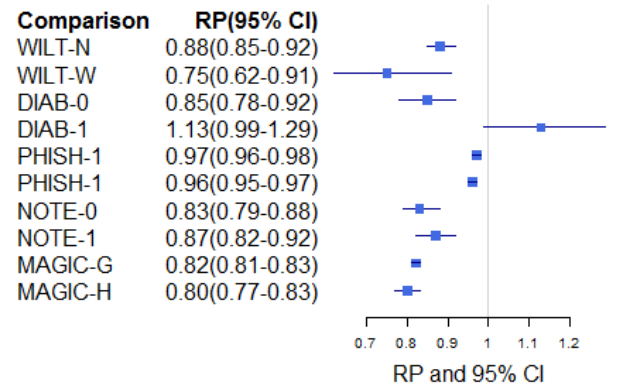


Figure 1: Forest plot for relative comparison with 95% confidence intervals

to derive new degrees of freedom, scaled test statistic in this case is an extension of (Lancaster 1961) and is given as

$$T_A = cT \approx \chi^2_\nu \quad (22)$$

where

$$c = \nu/E(T), \nu = 2(E(T)^2/var(T)) \quad (23)$$

and

$$E(T) = \sum_{i=1}^n w_i \quad (24)$$

$$Var(T) = 2 \sum_{i=1}^n w_i + 2 \sum_{i < j} p_{ij} \quad (25)$$

$$p_{ij} = \text{cov}(\gamma_{(w_i/2,2)}^{-1}(1 - p_i), \gamma_{(w_j/2,2)}^{-1}(1 - p_j)) \quad (26)$$

$T$  is test statistic from (Lancaster 1961) and  $p_{ij}$  takes correlated p-values into account. When the covariance  $p_{ij}$  is unknown, permutation or bootstrap methods can be used to simulate large enough sample (usually  $\geq 1000$ ) of p-values.

A much simpler method by Simes (Simes 1986) can be used as well. For an ordered set of  $L$  p-values, we have

$$\Pr \left\{ \bigcup_i^L (p_{(i)} < i\alpha/L) \right\} < \alpha \quad (27)$$

rejecting global  $H_0 = H_1, \dots, H_L$ , if  $p_{(i)} < i\alpha/L$  for at least one  $i$ . Global p-value is then given as  $\min\{L_{p(i)}/i\}$ . Originally designed for independent p-values, it has been shown (Sarkar and Chang 1997) to work well for positively correlated p-values.

To demonstrate, we focus on combining inference from one type of tests and on first five datasets where we would expect a combined test to reject global  $H_0$ . We used p-values from individual class comparisons using GS test. Method from (Hongying Dai and Cui 2014) was used to combine p-values as a positive correlation cannot be guaranteed. Table 7 shows the results. As expected, combined p-values are statistically significant at  $\alpha < 0.05$ . Rejecting global  $H_0$  in concordance with individual precision comparisons.

Table 7: Test of global null hypothesis using combined p-values

DATASET	CLASS	P-VALUE	COMBINED P-VALUE
WILT	N	<0.0001	<0.0001
	W	0.001	
DIABETES	0	0.0002	0.002
	1	0.07	
PHISHING	-1	<0.0001	<0.0001
	1	<0.0001	
BANK NOTE	0	<0.0001	<0.0001
	1	<0.0001	
MAGIC	G	<0.0001	<0.0001
	H	<0.0001	

## Multiple classifier comparison

GEE based marginal regression framework can be used to compare multiple multi-class classifiers. With a logistic regression model, using state-of-the-art classifier as the reference category, we can compare the performance of new proposed models to it. This procedure is valuable in large scale testing where we want to compare tens of classifiers to select a best fit for a problem. We have used two datasets to show its feasibility. Random forest with 50 trees (RF2) and Support Vector Machines (SVM) were added, resulting in a four classifier comparison. Results are shown in Table 8. P-values  $< \alpha$  inform us that at least one of the precision values is statistically significantly different from others. Magnitude

Table 8: Multiple classifier comparison

DATASET	CLASS	NB	RF1	SVM	RF2	P-VALUE
WILT	N	0.65	0.74	0.68	0.74	<0.0001
	W	0.73	0.98	1.0	0.95	<0.0001
DIABETIC	0	0.54	0.63	0.63	0.62	0.003
	1	0.76	0.67	0.72	0.66	0.12

Table 9: Multiple classifier comparison, odds ratio and 95% confidence intervals

DATASET	CLASS	COMPARISON	OR	LCL	UCL
WILT	N	RF1 vs NB	1.52	1.34	1.73
	N	SVM vs NB	1.12	1.03	1.22
	N	RF2 vs NB	1.51	1.33	1.71

OR: odds ratio, LCL: Lower confidence limit, UCL: Upper confidence limit

and size of difference can be estimated from parameter estimates/odds ratio and related CIs.

Table 9 shows the comparison of four learning algorithms on a class of dataset Wilt using odds ratio and related 95% confidence intervals. Only one class for one dataset is used for demonstration. Wealth of information provided by these estimates cannot be emphasized enough. Inferential statements such as "Random forest with 1000 trees has 52% (95% CI: 34%, 73%) higher chances of detecting non wilted trees compared to Naive Bayes" can be made. Odds ratio of greater than 1 confirms that model being compared is performing better than reference model.

## Partial Bayesian update of precision

We introduce here a special case of Bayes law, updating precision when class prevalence is known. As with most datasets used, it is understood that they are sampled from a larger population. When a class prevalence in population is known, precision can be updated as following

$$P_{Bayes} = \frac{S_s \times P_v}{S_s \times P_v + (1 - S_p) \times (1 - P_v)} \quad (28)$$

where  $S_s$  is classifier sensitivity,  $P_v$  is population prevalence and  $S_p$  is classifier specificity. This update is well known in medical statistics. Precision can be significantly changed with change in class prevalence (Altman and Bland 1994).

Classic example is from medical diagnostics where disease prevalence in a population is known and it can be used to update precision. Another example can be in object detection, as in indoor vs. outdoor images where prevalence of certain objects would be greater indoors (Tables, chairs, kettle etc.) and some outdoors (cars, buses, traffic signs etc.).

This update can be applied to most scenarios. Even using a justifiable assumption should yield better population level estimates compared to a non-informative approach. An example is shown in Table 10 where we have used diabetic retinopathy dataset to calculate precision. Then it is updated using population prevalence from (Schneider and Süveges

2004; Lee, Wong, and Sabanayagam 2015). Relative precision is the recommended method to compare updated precision values. Confidence intervals are suggested to be calculated using bootstrap methods. This update can still be used if prevalence is not known by substituting a normalized prevalence rate.

Table 10: Updating Precision using class prevalence

METHOD	CLASS	$P_{old}$	$P_{update}$
NB	0	0.54	0.35
RF	0	0.63	0.45
NB	1	0.76	0.87
RF	1	0.67	0.81

$P_{old}$  is empirical precision value and  $P_{update}$  is updated precision based on prevalence

### Comparison using Cross Validation

Methods described in this paper have been only applied in fixed train-test split. In this section we show their applicability when using cross validation. We used GEE based GLM on  $k$ -fold and  $n$  times repeated  $k$ -fold cross validation. For  $k$ -fold CV, we used a value of  $k = 10$ . For  $n$  times repeated  $k$ -fold CV, a value of 10 was used for  $n$  keeping  $k$  fixed at 10. After saving predictions for each fold and for each classifier, datasets are vertically stacked to generate a single dataset with multiple observations per record. Then a GEE based GLM is fitted. This is a slight modification to Algorithm 1. Results are reported in Table 11 using diabetic retinopathy dataset. Results from both cross validation variations agree with results using a fixed train-test split, albeit cross validation based statistical comparisons have more power in limited sample size setting. Same methods can be used for any resampling method used during CV.

Table 11: Results from 10 fold CV and  $10 \times 10$  fold CV

CLASS	10 FOLD CV	$10 \times 10$ FOLD CV
DISEASE	50.6 (<0.0001)	50.6 (<0.0001)
NON-DISEASE	10.2 (0.001)	10.2 (0.001)

Numbers outside parenthesis are test statistic with p-values inside

### Application to deep architectures

This section shows the application of precision based comparison to deep architectures. As new models are proposed often claiming to perform better than state-of-the-art, thorough comparisons are vital. We use two modified versions of deep convolution network described in (Simonyan and Zisserman 2014). For a simple demonstration we use the models to classify images of cats and dogs from Kaggle dataset (cats vs dogs) (Kaggle ). As the original model was trained on 1000 image classes including many instances of cats and dogs (Russakovsky et al. 2015), it has been modified to work

as a binary classifier (Chollet ). Two versions used differed only on dropout rate, first had a dropout rate of 0.5 and second 0.7. Precision outcomes for both classes from both models are reported in Table 12. As expected, overall accuracy of both models is very similar (90.76 and 90.88 respectively). But, the difference can be clearly seen in class breakdown where both models perform better on different classes. This type of analysis can also be used to adjust hyper-parameters for optimal performance.

Table 12: Comparing deep architectures using precision

CLASS	MODEL 1	MODEL 2	P-VALUE
DOGS	0.926	0.916	0.17
CATS	0.889	0.902	0.06

Values shown in table above are precision values compared using GS statistic

Although above described scenario is overly simplistic, it is to demonstrate the usefulness of presented methods to compare state-of-the-art. More complicated comparisons such as multiple object detection/classification in images can be implemented with similar ease.

### Conclusion and Recommendations

While machine learning literature is rich with evaluations and recommendations for statistical tests to compare classifiers based on classification accuracy, AUC, F-measure etc. It lacked a detailed study of statistical tests that can be used to compare classifiers based on precision or recall alone. Which are important performance metrics, especially for rare event classifiers. In this paper we have reviewed statistical methods based on marginal regression framework and Relative Precision. These can be used for classifier comparison using correlated precision values. We have presented empirical evaluation and implementation feasibility of these methods. As precision is usually calculated per-class, methods are presented to combine p-values for an overall classifier comparison. When a class prevalence is known, partial Bayesian update to precision is introduced. We have shown that the methods can be used in a cross validation setting and their application to compare deep architectures.

We recommend using GS statistic or RP for comparing two classifiers. Users concerned about use/misuse of p-values in statistical tests should use RP as results can be solely based on RP value and CIs. To simultaneously compare multiple classifiers, we recommend using GLM with GEE. Dai's method is recommended for combining dependent p-values over Simes' method as it retains appropriate power even when p-values are not positively correlated. Whenever possible, it is also recommended to use updated precision based on population prevalence.

### References

- [Altman and Bland 1994] Altman, D. G., and Bland, J. M. 1994. Statistics notes: Diagnostic tests 2: predictive values. *Bmj* 309(6947):102.

- [Anderson, Burnham, and Thompson 2000] Anderson, D. R.; Burnham, K. P.; and Thompson, W. L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *The journal of wildlife management* 912–923.
- [Aslan, Yıldız, and Alpaydın] Aslan, O.; Yıldız, O. T.; and Alpaydın, E. Statistical comparison of classifiers using area under the roc curve.
- [Benavoli et al. 2014] Benavoli, A.; Corani, G.; Mangili, F.; Zaffalon, M.; and Ruggeri, F. 2014. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 1026–1034.
- [Blake and Merz 1998] Blake, C., and Merz, C. J. 1998. {UCI} repository of machine learning databases.
- [Breiman 2001] Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.
- [Chollet] Chollet, F. Building powerful image classification models using very little data. (<https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html>. Accessed: 2016-07-29.
- [Demšar 2006] Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research* 7:1–30.
- [Dietterich 1998] Dietterich, T. G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation* 10(7):1895–1923.
- [Gill 1999] Gill, J. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52(3):647–674.
- [Halekoh, Højsgaard, and Yan 2006] Halekoh, U.; Højsgaard, S.; and Yan, J. 2006. s. *Journal of Statistical Software* 15(2):1–11.
- [Hongying Dai and Cui 2014] Hongying Dai, J., and Cui, Y. 2014. A modified generalized fisher method for combining probabilities from dependent tests. *Frontiers in genetics* 5.
- [Huber 1967] Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 221–233.
- [Joshi 2002] Joshi, M. V. 2002. On evaluating performance of classifiers for rare classes. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, 641–644. IEEE.
- [Kaggle] Kaggle. Kaggle dogs vs cats competition. (<https://www.kaggle.com/c/dogs-vs-cats>. Accessed: 2016-07-29.
- [Kosinski 2013] Kosinski, A. S. 2013. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in medicine* 32(6):964–977.
- [Lancaster 1961] Lancaster, H. 1961. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics* 3(1):20–33.
- [Lee, Wong, and Sabanayagam 2015] Lee, R.; Wong, T. Y.; and Sabanayagam, C. 2015. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye and Vision* 2(1):1.
- [Leisenring, Alono, and Pepe 2000] Leisenring, W.; Alono, T.; and Pepe, M. S. 2000. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* 56(2):345–351.
- [Li, Williams, and Cui 2011] Li, S.; Williams, B. L.; and Cui, Y. 2011. A combined p-value approach to infer pathway regulations in eqtl mapping. *Statistics and Its Interface* 4:389–401.
- [Liang and Zeger 1986] Liang, K.-Y., and Zeger, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 13–22.
- [McNemar 1947] McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12(2):153–157.
- [Nadeau and Bengio 2003] Nadeau, C., and Bengio, Y. 2003. Inference for the generalization error. *Machine Learning* 52(3):239–281.
- [Nickerson 2000] Nickerson, R. S. 2000. Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods* 5(2):241.
- [R Core Team 2015] R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [Russakovsky et al. 2015] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115(3):211–252.
- [Sarkar and Chang 1997] Sarkar, S. K., and Chang, C.-K. 1997. The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 92(440):1601–1608.
- [Satterthwaite 1946] Satterthwaite, F. E. 1946. An approximate distribution of estimates of variance components. *Biometrics bulletin* 2(6):110–114.
- [Schneider and Süveges 2004] Schneider, M., and Süveges, I. 2004. Retinopathia diabetica: magyarországi epidemiológiai adatok. *Szemészet* 141:441–444.
- [Simes 1986] Simes, R. J. 1986. An improved bonferroni procedure for multiple tests of significance. *Biometrika* 73(3):751–754.
- [Simonyan and Zisserman 2014] Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- [Stock and Hielscher 2013] Stock, C., and Hielscher, T. 2013. Dt-compair: comparison of binary diagnostic tests in a paired study design. *R package version 1*.
- [White 1980] White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society* 817–838.